COMPARING LANES IN THE PULSED-FIELD GEL ELECTROPHORESIS (PFGE) IMAGES

Wei-Zen Cheng [1] ,Kue-Sai Yen [1] ,Chih-Yang Lin [1] , Yu-Tai Ching [1] , Yun-Liang Yang [2]

[1] Department of Computer and Information Science, National Chiao Tung University, Hsin Chu Taiwan

[2] Department of Biological Science and Technology, National Chiao Tung University, Hsin Chu, Taiwan

Abstract- **Pulsed-field Gel Electrophoresis (PFGE) is an important tool in genomic analysis. The result of PFGE is presented in an image. Each image contains several lanes. And each lane consists of bands. Two lanes are identical if the relative positions of bands are the same. We present a method that uses computer to extract the lanes and compare the lanes in the electrophoresis images. The presented method consists of two major steps. The first step is image processing and lane extraction. The second step is to convert a lane into chain code representation. The lane comparison is carried out through calculating the longest common subsequence between lanes. We define the distance between lanes in term of the LCS and the lengths of two lanes. Two lanes have smaller distance tend to have similar pattern. This method eliminates those very different patterns to help biologists reduce the lanes that need to be compared.**

Keywords **- electrophoresis, image processing, chain code, dynamic programming.**

I. INTRODUCTION

Pulsed-field gel electrophoresis (PFGE) was developed in 1982 by Schwartz *et. al.* as a means of resolving very large DNA molecules [1]. PFGE can be used to separate DNA molecules from 10 kbp to approximately 10 Mbp and is an invaluable tool for genomic analysis. PFGE can be used for many applications in the different fields like biology, biochemistry, biotechnology, medicine, clinical diagnosis etc. This technique produces images that consist of several vertical lanes. Each lane contains some horizontal bands. Two subjects have the same gene sequence if their lanes have the same pattern. This work studied the method that uses computer techniques to identify the lanes and compare the lanes.

There are many factors that could affect the image quality, such as applied voltage and field strength, pulse time, reorientation angle, agarose type, concentration, the buffer chamber temperature, etc. [2-3]. Furthermore, the locations of the lanes and the size of the lanes in the image are different. All these factors make the lanes extraction and comparison difficult.

The images acquired in our system have grid-like texture in the background. The first step in the presented method is to remove the grid-like texture. The locations of lanes are than extracted by histogram analysis. The extracted lanes are then converted into chain code representation. Finally, comparison is carried out by calculating the LCS between lanes.

In the next section, we describe each step. The results are shown in Section 3.

II. METHOD

Preprocessing

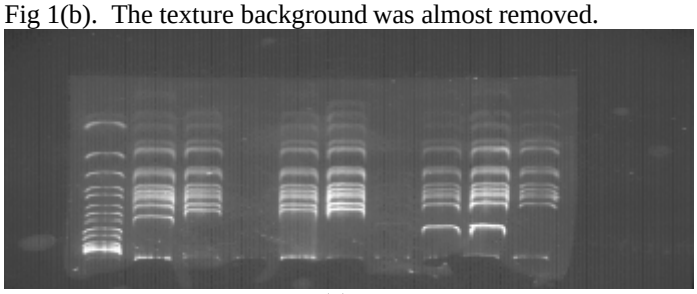
Let the added noise be denoted $n(x, y)$. Due to the added noise the PFGE images are blurry containing many grids and spot. We can express the PSE system as the following equation,

$$f(x,y) = \{O(x,y) * h(x,y) + n(x,y)\} * g(x,y)$$
$$= O(x,y) * h(x,y) * g(x,y) + n(x,y) * g(x,y). \quad (1)$$

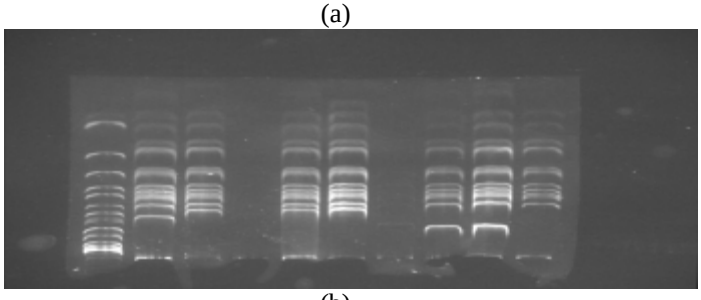
For the convenience of computing, we let the sampling frequency be 1. We apply the 1D Fourier Transform to $f(x, y)$ in the direction x.

$$F(x,y) = \sum_{k=0}^{N-1} f(x,y) e^{-j2pkx/N} \quad 0 \le x \le N-1, \quad (2)$$

We apply (1) for PGFE image (denoted $f(x, y)$). We found that there are some peaks of magnitude in certain frequency. We try to let $W(x, y)$ as a notch filter and apply it for Fig. 1(a) to remove the texture background. The result is shown in Fig 1(b). The texture background was almost removed.



(a)



(b)

Fig. 1. (a)Original image (b)After removing the texture background.

In order to remove the background, we apply histogram analysis of the images. Fig. 2(a), (b) show the histogram of before and after removing the texture background, respectively. In an image, the major part of the pixels belongs to the background that has lower intensity. After analyzing about 350 images, we noticed that the second peak

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Comparing Lanes in the Pulsed-Field GEL Electrophoresis (PFGE) Images	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Computer and Information Science National Chiao Tung University Hsin Chu Taiwan	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

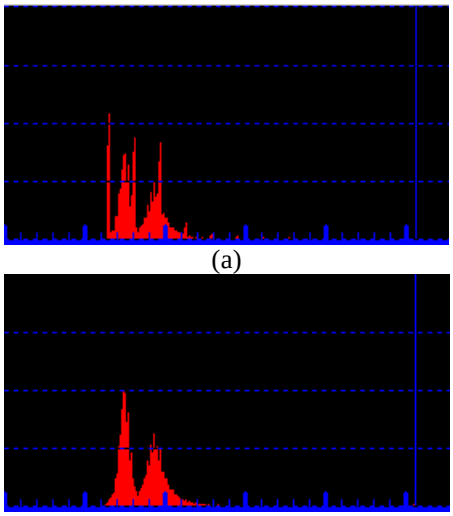
Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
3

in the histogram of an image can be set to the threshold to eliminate most of the background.



(a)

(b)

Fig. 2. The histogram of before (a) and after (b) removing the texture background.

Lanes Extraction

We assume that lanes are almost vertical. Given an image after preprocessing, we calculate the sum of the intensities on each column. A column in the background generally has low intensity. On the other hand, the accumulated intensities is high if a column passing through a lane. By examining the accumulated intensity, we can identify the location of the lanes. For each lane, we use the five columns in the center to represent the lane. Average intensities of rows along the column are converted to a polygonal curve as shown in Fig. 3.

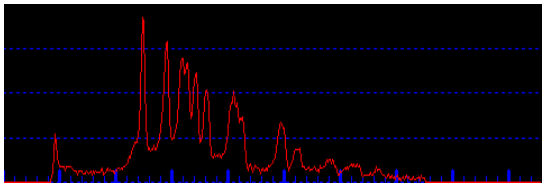


Fig. 3. the intensity distribution along the vertical line

In Fig. 3, the x-coordinate of the curve corresponds to the rows in the column and the y-coordinate represents the average intensity of rows. Averaging consecutive 5 rows along a column smoothes the polygonal path. The lanes and the associated polygonal paths are shown in Fig. 4.

Chain Code and Lane Comparison

Chain code encoding is one of the most fundamental techniques to present a polygonal path[4,5]. Chain code for a polygonal path is a stream of characters. The sequence of the characters describes the direction changes along the polygonal path. The chain code representation of a polygonal path provides an advantage that we can compare a pair of polygonal paths by comparing their chain code.

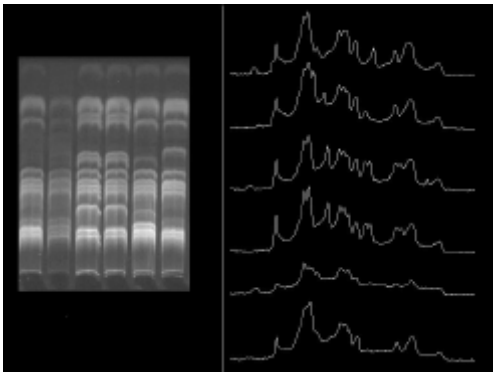


Fig. 4. Lanes and the associated polygonal paths.

Two polygonal paths are similar if their chain codes are similar. The similarity between two chain codes is defined by their longest common subsequence. Consider two chain codes s_1 and s_2 of lengths p and q, respectively. Let their longest common subsequence be denoted s, and the length of s is r. The distance between s_1 and s_2 is defined to be $(p+q-2r)/(p+q)$. Suppose s_1 and s_2 are identical, the distance between them is 0, otherwise the distance between the pair of chain codes is 1. The longest common subsequence of two chain codes can be obtained using the dynamic programming technique in $O(p*q)$ time.

The longest common subsequence of a pair of chain code can be represent pictorially as show in Fig. 5.

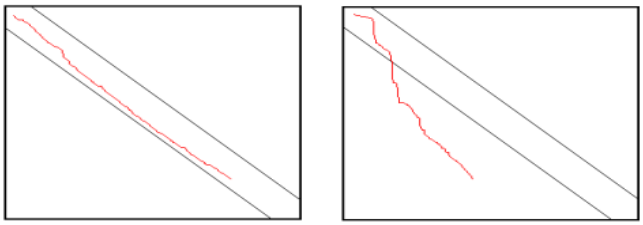


Fig. 5. The longest common subsequence of a pair of chain code.

Observe that, suppose s_1 and s_2 are identical, the common subsequence is shown as diagonal straight line. If s_1 and s_2 are very similar, the diagonal is a stair case curve close to the diagonal straight line. There are cases that the distance is short but the stair case line is far from the diagonal straight line. Furthermore, such lanes are perceptually different. That means, two lanes are similar if their chain codes meet the following conditions.

1. Distance between the pair of chain codes is short.
2. The stair case curve corresponding to the longest common subsequence is not far away from the diagonal.

III. Results

We applied the proposed method to compare a set of data consisting of 300 lanes. We first visually identified a pair of identical lanes. The threshold for the distance between and chain codes and the threshold for the distance between the stair case cure and the diagonal straight line are obtained from these pair of polygonal paths. We then computed the distances between all the pairs of the chain codes. All those

pairs whose distances are less than the given thresholds are possibly identical. And the determination of the similarity of these candidates is determined by the use of user assist system.

The presented method can serve as a screen system. Such screen system reduces tremendous amount of comparisons. In our experiment, there were 300 lanes so that there used to be 45,000 lanes. By using the proposed method, we can eliminate most of the pairs and leaves only 3000 candidates. This method could save many efforts for the biologist in comparing the lanes.

REFERENCES

[1] Schwartz, D.C., Saffran, W. Welsh, J., Haas, R., Goldenberg, M., and Cantor, C.R. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 198-195

[2] Van Daelen, R.A.J., and Zabel, P. (1991). Preparation of high molecular weight plant DNA and analysis by pulsed-field gel electrophoresis. In *Plant Molecular Biology Manual* (S.B. Gelvin, R.A. Schilperoort, and D.P.S. Verma, eds.), pp. A15/1-25. Kluwer Academic Publishers, The Netherlands.

[3] Birren, B., Hood, L., and Lai, E. (1989). "Pulsed field gel electrophoresis: Studies of DNA migration made with the programmable, autonomously-controlled electrode electrophoresis system." *Electrophoresis* 10, pp. 302-309.

[4]Haim J. Wolfson, "On Curve Matching" , IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12 ,No.5 ,MAY 1990

[5] Ernesto Bribiesca, "A New Chain Code", Pattern Recognition 32 ,p.p235-251,1999